Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

Комп’ютерний практикум №1

«Експериментальна оцінка ентропії

на символ джерела відкритого тексту»

Виконали

студенти групи ФБ-93

Бурячок А.А

Данілін Д.Д.

Перевірила

Селюх П.В.

Київ - 2021

Мета: вивчення поняття ентропії, її експериментальна оцінка. Дослідження понять ентропії на символ джерела та його надлишковості. Набуття практичних навичок з програмування, оцінки ентропії на символ джерела та надлишковості тексту.

Завдання:

- уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- написати програму для підрахунку частот літер і частот біграм в тексті, а також підрахунку H_1, H_2 за безпосереднім означенням. Підрахувати частоти літер та біграм, а також значення H_1, H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1 Мб), де ймовірності заміняються відповідними частотами. Також отримати значення H_1, H_2 на тому ж самому тексті, в якому вилучено всі пробіли.
- за допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30).
- використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

Частина 1

Пишемо програму, яка буде рахувати частоти літер і біграм в тексті, а також обчислювати значення H_1 та H_2 за безпосереднім означенням. Для обчислення частот літер у тексті ми використали модуль "Counter". Він дозволяє отримати частоту літер для деякого рядка або тексту. Щодо обчислення частот біграм, то там не все так просто. Для підрахунку біграм, що не перетинаються ми використали цикл, який ітерується по всьому тексту з кроком 2 і додає до відповідного лічильника біграми, яку ми отримали, одиницю. Щодо перехресних біграм, то там все так само, але ітеруємося з кроком 1. Код програми наведено нижче:

У результаті ми отримуємо частоти літер та біграм у тексті (результати наведені у файлах у папці results).

Частоти літер.

З пробілами	Без пробілів

" " => 0.1693387615987945 о => 0.09350257319831655 е => 0.07423390154424236 а => 0.06606320868048202 н => 0.05567259669097853 и => 0.05300885542392694 т => 0.052353227451975855 с => 0.044177684073009954 в => 0.039139615612838345 л => 0.038553511692166906 р => 0.03242407731074502 к => 0.026989075022918686 м => 0.02596965841055084 д => 0.025518560496434063 п => 0.022496689523372207 у => 0.021803065848977593 я => 0.019828906160315993 ь => 0.01896389761532504 ч => 0.01586765207577798 г => 0.015596023224266804 з => 0.014414922771713735 ы => 0.0144076469989054 б => 0.013902384998326573 ж => 0.009792381780818152 й => 0.008166650767755716 ш => 0.006784253934172041 4 х => 0.006044550365324637 ю => 0.004969352828092891 э => 0.002970132144202583 4 щ => 0.002444659663600602 ц => 0.0024082807995589265 ф => 0.0015174028401383367	о => 0.11257943976215803 е => 0.09018336898624928 а => 0.07954594160714773 н => 0.06703819732622612 и => 0.06381969365767745 т => 0.06302383485101756 с => 0.05317934390841682 в => 0.047131985925865316 л => 0.04641892760704642 р => 0.03899610740294809 к => 0.03249773029250977 м => 0.031251826481349436 д => 0.03072677397609891 п => 0.02708647488125436 у => 0.026243858152048972 я => 0.023866022965932956 ь => 0.022830529806227375 ч => 0.01912399032111003 г => 0.018778176518950595 з => 0.017357904621630994 ы => 0.017350111634540078 б => 0.01673349153097128 ж => 0.011788841221784516 й => 0.009827930845032555 ш => 0.008163153977735435 х => 0.007273779325984547 ю => 0.005978195222119615 э => 0.003576006951344485 щ => 0.0029447749969802174 ц => 0.0028882758405710703 ф => 0.001795309401069977

В таблиці нижче наведені перші 10 найчастіших біграм

Частоти біграм з кроком 1 без пробілів	Частоти біграм з кроком 1 з пробілами	Частоти біграм з кроком 2 без пробілів	Частоти біграм з кроком 2 з пробілами
то : 0.017742 не : 0.013435 но : 0.012735 на : 0.011974 ст : 0.011974 ов : 0.011765 ен : 0.010954 по : 0.010604	о_ : 0.025275 е_: 0.018942 и_: 0.017320 _в : 0.016908 _н : 0.016763 а_ : 0.016091 _с : 0.015478 _п : 0.014957	то : 0.018051 не : 0.013462 но : 0.012706 на : 0.012003 ст : 0.012001 ов : 0.011820 ен : 0.011167 по : 0.010485	о _: 0.025431 е_ : 0.018912 и _: 0.017347 _в : 0.016899 _н : 0.016624 а _: 0.015999 _с : 0.015517 _п : 0.015168

го : 0.010143 ко : 0.009559	то : 0.014465 ь_ : 0.012505	го : 0.010203 ко : 0.009505	то : 0.014218 ь_ : 0.012377

Після того як ми підрахували частоти літер і біграм у тексті, програма обчислює значення ентропії. За означенням вона з точністю до знаку дорівнює сумі добутків ймовірності на її логарифм. У нашому випадку, за законом великих чисел, ймовірність дорівнює частоті літери або біграми. Ми отримали наступні значення ентропії:

Текст з пробілами:

Entropy for letters: 4.36270652063231; $R = 12.7\%$

Entropy for bigrams with step 1: 3.9432236058456143;$R = 21.1\%$

Entropy for bigrams with step 2: 3.9435518502674785;$R = 21.1\%$

Якщо з тексту видалити пробіли, то отримали наступні частоти літері біграм (результати наведені у файлах у папці results).

Відповідно трохи змінилися значення ентропії (результати наведені нижче).

Текст без пробілів:

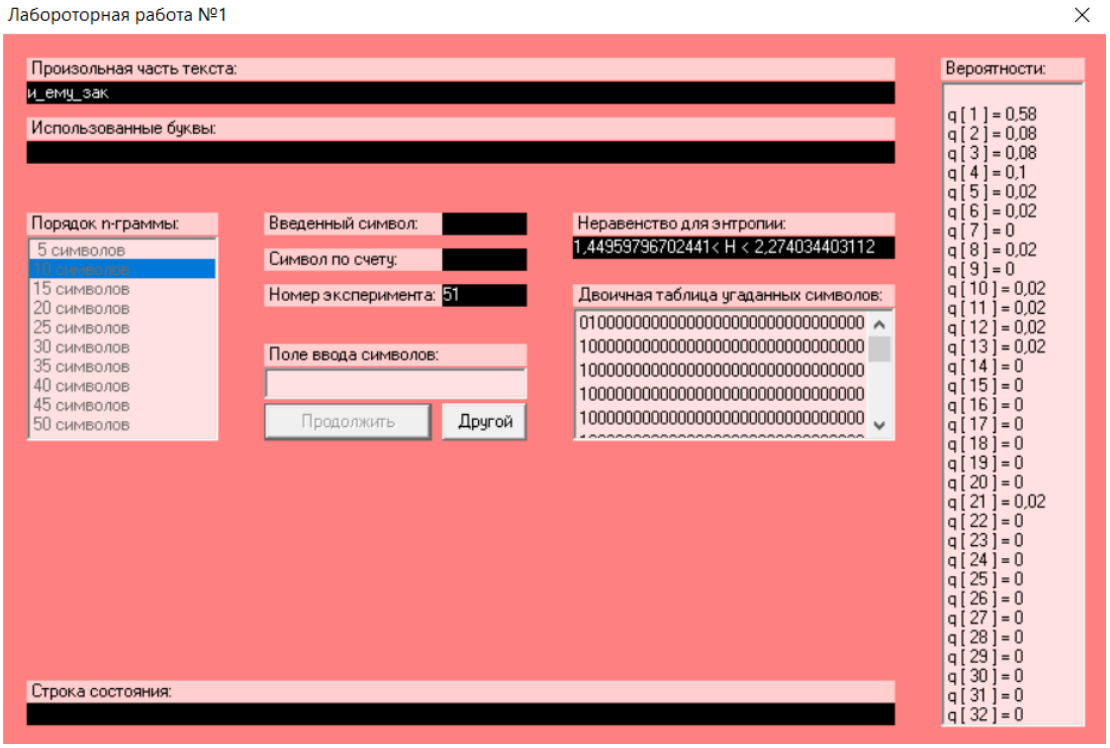
Entropy for letters: 4.455195190668355;$R = 10.7\%$

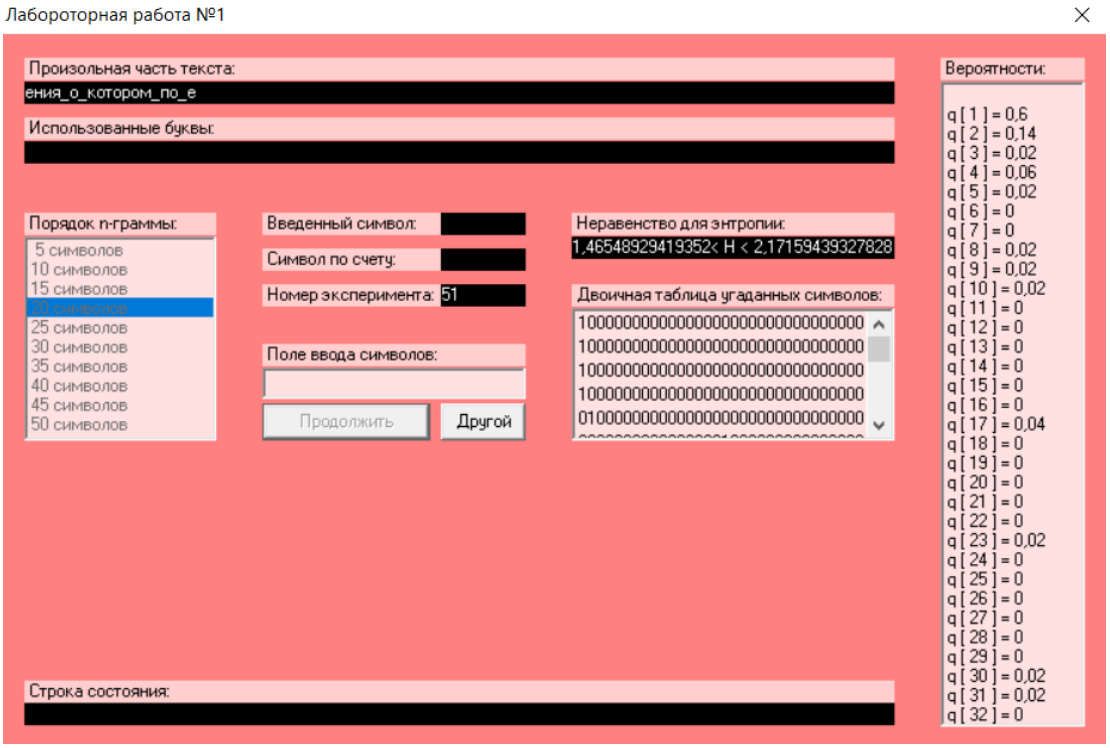
Entropy for bigrams with step 1: 4.127174564393386;$R = 17.4\%$

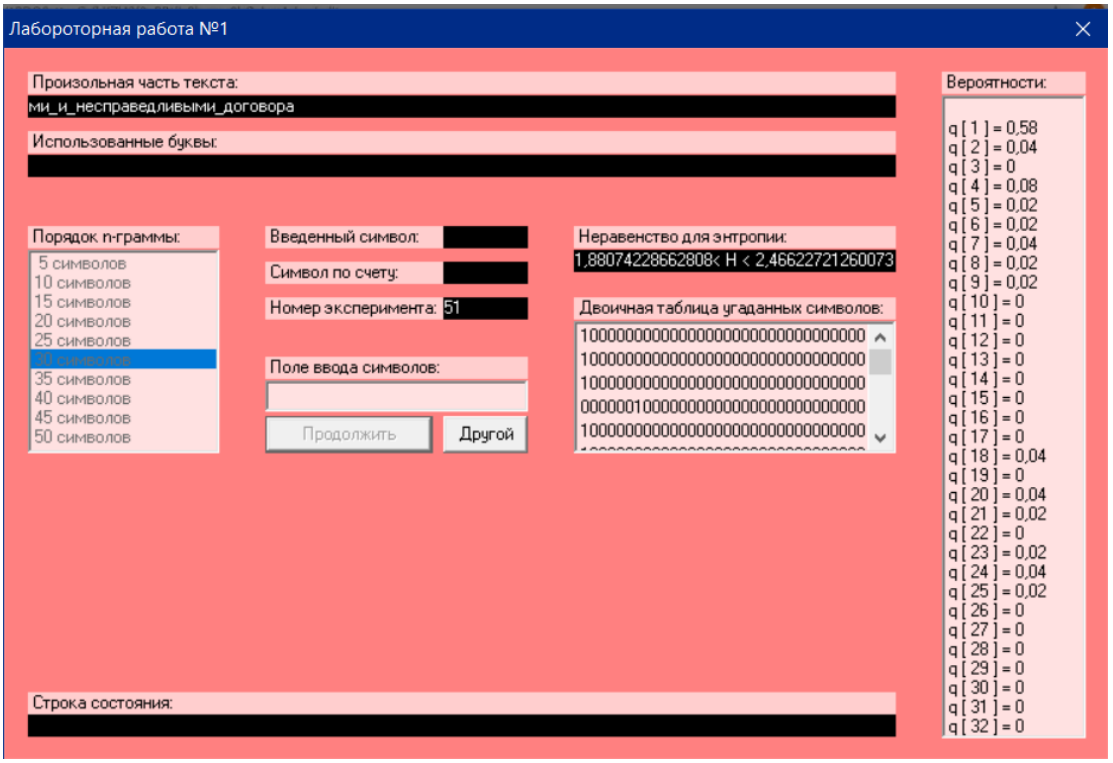
Entropy for bigrams with step 2:4.126521022690573;$R = 17.5\%$

Частина 2

Запускаємо CoolPinkProgram, у якій нам необхідно, використовуючи частину тексту, вгадати яким буде наступний символ і на основі цих даних оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$та надлишковості російської мови в різних моделях джерел.


Лабороторная работа №1
Произольная часть текста:
и_ему_зак
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,44959796702441< H < 2,274034403112
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,58
q [2] = 0,08
q [3] = 0,08
q [4] = 0,1
q [5] = 0,02
q [6] = 0,02
q [7] = 0
q [8] = 0,02
q [9] = 0
q [10] = 0,02
q [11] = 0,02
q [12] = 0,02
q [13] = 0,02
q [14] = 0
q [15] = 0
q [16] = 0
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0,02
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0


Лабороторная работа №1
Произольная часть текста:
ения_о_котором_по_е
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,46548929419352< H < 2,17159439327828
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,6
q [2] = 0,14
q [3] = 0,02
q [4] = 0,06
q [5] = 0,02
q [6] = 0
q [7] = 0
q [8] = 0,02
q [9] = 0,02
q [10] = 0,02
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0
q [17] = 0,04
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0,02
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0,02
q [31] = 0,02
q [32] = 0



Отримали наступні значення ентропії:

$$H^{(10)} = 1.86, H^{(20)} = 1.81, H^{(30)} = 2.17.$$

Обчислюємо значення надлишковості джерела відкритого тексту:

$$R_1 = 1 - \frac{1.86}{5} = 0.628,$$

$$R_2 = 1 - \frac{1.81}{5} = 0.638,$$

$$R_3 = 1 - \frac{2.17}{5} = 0.566.$$

Отже, надлишковість російської мови становить 62.8%, 63.8% і 56.6% в проведених експериментах.

Висновки: в ході лабораторної роботи ми вивчили поняття ентропії на символі джерела та його надлишковість, дослідили та порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії, а також набули практичних навичок щодо оцінки ентропії на символі джерела та вдосконалили знання у сфері програмування. Під час аналізу тексту ми підтвердили той факт, що в російському

алфавіті частіше всього зустрічається пробіл, що свідчить про те, що при шифруванні потрібного його прибирати, щоб зловмиснику требу було прикласти більше зусиль для його зламу. Якщо з тексту його прибрати, то це будуть літери “о”, “е”, “а”, а рідше за все зустрічаються: “ф”, “ц”. Серед біграм у тексті з пробілом частіше всього зустрічаються: “о_”, “е_”, “и_”, а у тексті без пробілів: “то”, “не”, “но”. За допомогою CoolPinkProgram отримали що надлишковість в середньому по трьом дослідам російської мови 61%.